

# The two-star model: exact solution in the sparse regime and condensation transition

A Annibale<sup>†</sup>, OT Courtney

<sup>†</sup> Department of Mathematics, King’s College London, The Strand, London WC2R 2LS, UK

**Abstract.** The 2-star model is the simplest exponential random graph model that displays complex behavior, such as degeneracy and phase transition. Despite its importance, this model has been solved only in the regime of dense connectivity. In this work we solve the model in the finite connectivity regime, far more prevalent in real world networks. We show that the model undergoes a condensation transition from a liquid to a condensate phase along the critical line corresponding, in the ensemble parameters space, to the Erdős-Rényi graphs. In the fluid phase the model can produce graphs with a narrow degree statistics, ranging from regular to Erdős-Rényi graphs, while in the condensed phase, the “excess” degree heterogeneity condenses on a single site with degree  $\sim \sqrt{N}$ . This shows the unsuitability of the two-star model, in its standard definition, to produce arbitrary finitely connected graphs with degree heterogeneity higher than Erdős-Rényi graphs and suggests that non-pathological variants of this model may be attained by softening the global constraint on the two-stars, while keeping the number of links hardly constrained.

## 1. Introduction

Exponential random graph models (ERGM) are ensembles of random graphs where each graph configuration  $\mathbf{c}$  appears with a probability  $p(\mathbf{c}) \propto e^{-H(\mathbf{c})}$  given by the Gibbs-Boltzmann distribution, where  $H(\mathbf{c})$  is the graph Hamiltonian, enclosing several properties of the networks in the ensemble. First introduced in the 1980s by Holland and Leinhardt [22], and further developed by Frank and Strauss [19] and in several later studies [7, 24, 3, 6, 27], ERGM soon became popular in social network analysis. Computational tools to analyze and simulate networks based on ERGM are largely available on the web, as the ERGM and SIENA packages, and several paradigmatic models of random graphs can be written in the exponential form, for suitable choices of the graph Hamiltonian, including the Erdős-Rényi (ER) graph ensemble [20] and the ensemble of graphs with soft-constrained degree sequence [8, 4, 16]. However, ERGM have known drawbacks which limit their practical use as proxies or null models for real networks. In particular, they may display degeneracy behaviour and may fail to produce graphs with properties within certain ranges, which are nevertheless observed in nature.

A useful model to understand degeneracy behavior and limitations of ERGM, is the two-star model, which is simple enough to be amenable of exact solutions, while exhibiting the interesting behavior of more complex ERGM. A well-known solution for the two-star model has been developed, in the limit of large system size  $N$ , by Park and Newmann [12], within a mean-field approach and expansions around it. However, this approach requires that the connectivity

of the graph is  $\mathcal{O}(N)$ , a condition which is hardly met by real world networks, like social and biological networks, where the connectivity is  $\mathcal{O}(N^0)$ . In this work we solve the two-star model exactly, in the finite connectivity regime, for large system size, by using path integrals. We check our calculations against Monte-Carlo simulations and compare our results with those predicted by mean-field theories. Results show that the 2-star model undergoes a condensation transition from a liquid to a condensate state, along the critical line corresponding, in the ensemble parameters space, to the Erdős-Rényi graphs. In the liquid phase the model can only produce graphs within a narrow range of degree statistics, namely between regular and Erdős-Rényi (ER), whereas in the condensed phase a condensate of size  $\sim \sqrt{N}$  emerges, residing on a single site.

The paper is structured as follows: in Sec. 2 we define the model and review the mean-field solution, in Sec. 3 we solve the two-star model in the finite connectivity regime and compare our results with the mean-field predictions and with Monte-Carlo simulations. In Sec. 4 we show that the phase transition displayed by the system in the finite connectivity regime is a condensation, related to large fluctuations in the sums of random variables. Finally, we summarise our conclusions in Sec. 5.

## 2. The model and the mean-field solution

The 2-star model is a prototypical example of ERGM. In this section, we give a brief review of ERGM, the 2-star model and its mean-field solution.

### 2.1. Background: the ERGM

For simple and undirected graphs of  $N$  nodes, ERGM are graph ensembles where each graph is defined by an adjacency matrix  $\mathbf{c}$ , with elements  $c_{ij} \in \{0, 1\} \forall i, j = 1, \dots, N$ ,  $c_{ii} = 0 \forall i$ ,  $c_{ij} = c_{ji} \forall i < j$ . In ERGM, each graph  $\mathbf{c}$  appears with a probability given by the Gibbs-Boltzmann distribution

$$p(\mathbf{c}) = \frac{1}{Z} e^{-H(\mathbf{c})} \quad (1)$$

with Hamiltonian  $H(\mathbf{c})$  and partition function  $Z = \sum_{\mathbf{c}} e^{-H(\mathbf{c})}$ . The Hamiltonian is given by

$$H(\mathbf{c}) = - \sum_{\mu=1}^K \lambda_{\mu} \Omega_{\mu}(\mathbf{c}) \quad (2)$$

where  $\boldsymbol{\Omega}(\mathbf{c}) = (\Omega_1(\mathbf{c}), \dots, \Omega_K(\mathbf{c}))$  is a set of observables of the graph  $\mathbf{c}$  for which one has statistical estimates  $\boldsymbol{\Omega} = (\Omega_1, \dots, \Omega_K)$ , so that

$$\langle \Omega_{\mu}(\mathbf{c}) \rangle = \Omega_{\mu} \quad \forall \mu = 1, \dots, K \quad (3)$$

with  $\langle \Omega_{\mu}(\mathbf{c}) \rangle = \sum_{\mathbf{c}} p(\mathbf{c}) \Omega_{\mu}(\mathbf{c})$  and  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$  are ensemble parameters also called “conjugated” observables, which have to be calculated from the equations for the constraints (3). The distribution (1) with Hamiltonian (2) maximises the Shannon entropy

$$\mathcal{S}(\boldsymbol{\Omega}) = - \sum_{\mathbf{c}} p(\mathbf{c}) \ln p(\mathbf{c}) \quad (4)$$

subject to the constraints (3) and the normalization  $\sum_{\mathbf{c}} p(\mathbf{c}) = 1$ . Hence, ERGM are maximum entropy ensembles conditioned on the imposed constraints.

For all but the simplest ERGM, exact solutions for the ensemble parameters are not available, so one has normally to take recourse to numerical methods. The latter have been the subject of intense investigation over the last few decades and range from pseudo-likelihood estimation [15, 7, 5, 21, 24, 23] to Markov Chain Monte Carlo Maximum Likelihood techniques [28, 17, 26, 23], to Bayesian inference [10, 1, 2].

Once the values of the parameters  $\boldsymbol{\lambda}$  are available, one can use the resulting probability distribution  $p(\mathbf{c})$  to estimate the value of observables for which no estimate is available. We note that expectation values for the primary network observables  $\{\Omega_\mu\}$  are given by

$$\langle \Omega_\mu(\mathbf{c}) \rangle = \frac{1}{Z(\boldsymbol{\lambda})} \sum_{\mathbf{c}} \Omega(\mathbf{c}) e^{\sum_\mu \lambda_\mu \Omega_\mu(\mathbf{c})} = -\frac{\partial F(\boldsymbol{\lambda})}{\partial \lambda_\mu} \quad (5)$$

where  $F(\boldsymbol{\lambda}) = -\ln Z(\boldsymbol{\lambda})$  is the free-energy, and their fluctuations are given by the second derivative of the free-energy  $\langle \Omega_\mu^2 \rangle - \langle \Omega_\mu \rangle^2 = \partial^2 F / \partial \lambda_\mu^2$  which, in equilibrium, equates the so-called susceptibility, defined as  $\chi_\mu = \partial \langle \Omega_\mu \rangle / \partial \lambda_\mu$ , and measuring the deviation in  $\langle \Omega_\mu \rangle$  when a change in the external variable  $\lambda_\mu$  is applied. These relations suggest that in general we are able to calculate the ensemble parameters from the equations for the constraints, if we can calculate the free energy of the ERGM. This is straightforward only for Hamiltonians which are linear in the adjacency matrix, where the graph distribution factorizes over the links  $p(\mathbf{c}) = \prod_{i < j} p(c_{ij})$ . For non-linear Hamiltonians, like the two-star model and Strauss model [7], solutions have been found within mean-field approximations and expansions about it [12, 14].

## 2.2. The 2-star model

In the two-star model, one chooses, as graph observables, the number of links

$$L(\mathbf{c}) = \frac{1}{2} \sum_{i \neq j} c_{ij} \quad (6)$$

and the number of two-stars (i.e. paths of length two)

$$S(\mathbf{c}) = \frac{1}{2} \sum_{i \neq j \neq k (\neq i)} c_{ij} c_{jk} \quad (7)$$

and assumes that the expectations  $L = \langle L(\mathbf{c}) \rangle$  and  $S = \langle S(\mathbf{c}) \rangle$  are known. This leads to the non-linear network Hamiltonian

$$\begin{aligned} H(\mathbf{c}) &= -\theta_1 L(\mathbf{c}) - \theta_2 S(\mathbf{c}) = -\sum_{i < j} c_{ij} (\theta_1 + \theta_2 \sum_{k \neq i, j} c_{jk}) \\ &= -\sum_{i < j} c_{ij} \left[ (\theta_1 - \theta_2) + \theta_2 \sum_k c_{jk} \right] \\ &= -\sum_{i < j} c_{ij} \left( 2\alpha + 2\beta \sum_k c_{jk} \right) \end{aligned} \quad (8)$$

where we set  $\alpha = (\theta_1 - \theta_2)/2$  and  $\beta = \theta_2/2$ . Alternatively, we can define the local degrees of graph  $\mathbf{c}$  as  $k_i(\mathbf{c}) = \sum_j c_{ij} \forall i$ , and write the observables (6) and (7) as

$$\begin{aligned} L(\mathbf{c}) &= \frac{1}{2} \sum_i k_i(\mathbf{c}) \\ S(\mathbf{c}) &= \frac{1}{2} (\sum_j k_j^2(\mathbf{c}) - \sum_j k_j(\mathbf{c})) \end{aligned} \quad (9)$$

so that the Hamiltonian reads

$$H(\mathbf{c}) = -\alpha \sum_i k_i(\mathbf{c}) - \beta \sum_j k_j^2(\mathbf{c}). \quad (10)$$

At this point it is useful to define the average connectivity  $\bar{k}(\mathbf{c}) = N^{-1} \sum_i k_i(\mathbf{c})$  and the expected average connectivity in the ensemble  $\langle k \rangle = \langle \bar{k}(\mathbf{c}) \rangle$ , with  $\langle \cdot \rangle$  denoting the average over the ensemble probability  $p(\mathbf{c})$ . In the two-star model, we have  $L = N\langle k \rangle/2$  and  $S = N(\langle k^2 \rangle - \langle k \rangle)/2$ , hence we constrain  $\langle k \rangle$  and  $\langle k^2 \rangle$ .

The two-star model has so far been solved for dense graphs, with average connectivity  $\langle k \rangle = \mathcal{O}(N)$ , for which mean-field approaches are exact in the thermodynamic limit. The mean-field solution (see Section 2.3) shows that for large  $N$ , and  $\beta N = 2J$ , the two-star model undergoes, for any  $J \geq 1$ , a first-order phase transition, between a phase of low connectivity and one of high connectivity, separated by the critical line  $J = -\alpha$  in the space of ensemble parameters. The critical line terminates at the critical point  $J = 1$ , where a second order phase transition takes place from the symmetry-broken state with two phases, the one with high and the one with low connectivity respectively, to a symmetric state [12, 9]. In particular, for  $J \geq 1$  the link density is discontinuous meaning that the model fails to produce intermediate connectivities by suitably tuning the ensemble parameters. Along the critical line  $J = -\alpha$  (and for  $J \geq 1$ ) one has degeneracy, meaning that for the same ensemble parameters the model produces either a sparse or a dense graph.

However, it is not clear a priori how well the mean-field scenario applies to the finite connectivity regime, where Gaussian fluctuations (around the mean-field solution) become dominant rather than a small perturbation around the leading order statistics. In addition, we are interested in establishing whether in the phase at low connectivity the model can produce arbitrary densities of stars for any given (finite) connectivity, by choosing suitably the ensemble parameters. More in general, we aim to establish whether the 2-star model may serve as a plausible null-model for finitely connected networks, which are far more prevalent than dense networks in the real world. We answer these questions precisely in Section 3 by solving the two-star model exactly, in the limit  $N \rightarrow \infty$ , for finite connectivity  $\langle k \rangle = \mathcal{O}(N^0)$ .

### 2.3. The mean-field approach

In this section we briefly illustrate the mean-field approach, that is exact, in the thermodynamic limit, for dense graphs [13], but it is expected to give inaccurate results for finite connectivity. By analogy with spin models, we can regard the expression in the brackets of (8) as the local field acting on the link  $c_{ij}$ . The mean-field approximation replaces the local fields with their ensemble

averages, i.e.  $c_{jk} \rightarrow \langle c_{jk} \rangle$ . Doing so, all edges in the model become equivalent and the average probability to observe a link is the same for all links  $p = \langle c_{jk} \rangle \forall j, k$ . Hence, the Hamiltonian becomes

$$H(\mathbf{c}) = -\lambda \sum_{i < j} c_{ij} \quad (11)$$

where

$$\lambda = 2\alpha + 2\beta(N-1)p \simeq 2\alpha + 2\beta Np \quad (12)$$

is now a function of the unknown probability  $p$  and the last approximation holds for  $N \gg 1$ . The Hamiltonian (11) leads to the partition function

$$Z = \sum_{\mathbf{c}} e^{\lambda \sum_{i < j} c_{ij}} = \prod_{i < j} \sum_{c_{ij}} e^{\lambda c_{ij}} = \prod_{i < j} (1 + e^{\lambda}) = (1 + e^{\lambda})^{N(N-1)/2} \quad (13)$$

which immediately gives the free energy

$$F = -\ln Z = -\frac{N(N-1)}{2} \ln(1 + e^{\lambda}) \quad (14)$$

that can be used to write the equation for the constraint

$$L = -\frac{\partial F}{\partial \lambda} = \frac{N(N-1)}{2} \frac{e^{\lambda}}{1 + e^{\lambda}}. \quad (15)$$

If links are all drawn randomly and independently with probability  $p$ , we have

$$L = \frac{N(N-1)}{2} p \quad (16)$$

hence we get from (15)

$$p = \frac{e^{\lambda}}{1 + e^{\lambda}} = \frac{1}{2} \left( 1 + \tanh \frac{\lambda}{2} \right) \quad (17)$$

where in the last equality we used the identity  $e^x/2 \cosh x = (1 + \tanh x)/2$ . Now, however,  $\lambda$  is a function of  $p$ , so the above gives a self-consistency equation for  $p$

$$p = \frac{1}{2} [\tanh(\beta Np + \alpha) + 1] \quad (18)$$

This equation is identical to the one found by Park and Newman [13] by solving the model exactly in the dense regime, i.e. for  $p = \mathcal{O}(1)$ , and by setting  $\beta N = 2J$

$$p = \frac{1}{2} [\tanh(2Jp + \alpha) + 1] \quad (19)$$

As noted in [13], a unique solution to (19) exists only for  $J < 1$ . For  $J \geq 1$  and  $\alpha$  sufficiently close to  $-J$  there are three solutions, with only the outer two being stable, leading to a degeneracy in the solution. Expansions around the mean-field solution and perturbation theories [12, 14] give for the first two moments of the degree distribution

$$\langle k \rangle = Np + \frac{4Jp(1-p)(1-2p)}{[1 - 8Jp(1-p)][1 - 4Jp(1-p)]} \quad (20)$$

$$\langle k^2 \rangle = N^2 p^2 + \frac{Np(1-p)(1-8Jp^2)}{[1 - 8Jp(1-p)][1 - 4Jp(1-p)]} \quad (21)$$

where the second terms on the RHS, originate from the Gaussian fluctuations about the mean-field solution. These are subleading for large  $N$ , in the dense regime where  $p = \mathcal{O}(1)$  and  $J = \mathcal{O}(1)$ . However, in the finite connectivity regime where  $p = \mathcal{O}(N^{-1})$  and  $\beta = \mathcal{O}(1)$  (see Appendix A), Gaussian fluctuations are no longer small fluctuations about the leading order statistics. Upon setting  $p = c/N$  and sending  $N \rightarrow \infty$  at constant  $\beta$  and  $c$ , we get

$$\langle k \rangle = c \left[ 1 + \frac{\beta}{(1 - 2\beta c)(1 - \beta c)} \right] \quad (22)$$

$$\langle k^2 \rangle = c \left[ c + \frac{1}{(1 - 2\beta c)(1 - \beta c)} \right] \quad (23)$$

showing that the mean-field approximation becomes inexact in the finite connectivity regime. For a full discussion of mean-field predictions in the finite connectivity regime see Appendix A.

#### 2.4. Upper and lower bounds on the total number of stars in the finite connectivity regime

Before solving the model in the finite connectivity regime, it is useful to derive expressions for the upper and lower physical bounds on the number of stars that the model can exhibit at a given finite connectivity. The expected number of 2-stars is  $S = N(\langle k^2 \rangle - \langle k \rangle)/2$ . For a fixed number of edges  $L$ , the total number of stars is minimised by minimising the second moments of the degree distribution while keeping the first moment fixed, i.e. by making the degree distribution regular, so that  $\langle k^2 \rangle = \langle k \rangle^2$  attains its physical minimum. This corresponds to the minimum star density

$$\frac{S_{\min}}{N} = \frac{1}{2} \langle k \rangle (\langle k \rangle - 1) \quad (24)$$

The total number of stars is maximised by maximising the second moment while keeping the first moment fixed, resulting in a small number of vertices having very large degrees while all the others have low degrees. To see this, we consider the following iterative process. We pick two vertices  $i$  and  $j$  and assume that  $k_i \geq k_j$ . If  $j$  has a neighbour which is not already connected to  $i$  then this edge is rewired to increase  $k_i$  by 1 and decrease  $k_j$  by 1. This step is repeated until in every pair, the vertex with the lesser degree has no more 'spare' edges to rewire to the other vertex. In the case where  $L \leq N - 1$  this always ends with a graph where there is one vertex with degree  $L$ ,  $L$  vertices connected to it with degree 1, and any left over vertices having degree 0 (because any other configuration would contain 'spare' edges). If we increase  $L$  just beyond  $N$ , we can no longer increase the stars by rewiring to the original hub as that hub is already connected to every other vertex. This causes a new hub to form to take up the extra edges. This process of hubs being born and increasing in size until they have degree  $N$  keeps going as  $L$  is increased until the graph is complete.

In this work we consider sparse graphs, where the average connectivity is  $\mathcal{O}(1)$ . For this case the total number of edges  $L$  is  $\mathcal{O}(N)$  which means that the minimum value that the number of stars can take is  $\mathcal{O}(N)$ . On the other hand, the maximum star configuration has an  $\mathcal{O}(1)$  number of vertices with degree  $N$ , and an  $\mathcal{O}(N)$  number of vertices with degree  $\mathcal{O}(1)$ . The vertices with degree  $N$  give each a contribution to the number of stars  $N(N - 1)/2$  showing that the maximum density of stars in a finitely connected graph is

$$\frac{S_{\max}}{N} \sim \mathcal{O}(N). \quad (25)$$

Expansions about mean-field solutions (22), (23) show that both moments  $\langle k \rangle$ ,  $\langle k^2 \rangle$  diverge at the same parameter values, suggesting that it is impossible to achieve the maximum expected star density for any finite value of connectivity. However, higher order (non-Gaussian) fluctuations may become dominant in the finite connectivity regime and the expansions (22) and (23) may get very inaccurate. We will test their accuracy against MCMC simulations and results of the exact calculation in the next section.

### 3. Exact solution in the finite connectivity regime

In this section, we derive an exact solution for the two-star model in the finite connectivity regime, where mean-field solutions are expected to become inexact. First off, we rewrite the graph Hamiltonian by performing the sum over  $k$  in (8) and using symmetry of  $c_{ij} = c_{ji}$

$$H(\mathbf{c}) = -2\alpha \sum_{i < j} c_{ij} - \beta \sum_{i < j} c_{ij} (k_i(\mathbf{c}) + k_j(\mathbf{c})). \quad (26)$$

Hence, we introduce the partition function

$$\begin{aligned} Z &= \sum_{\mathbf{c}} e^{2\alpha \sum_{i < j} c_{ij} + \beta \sum_{i < j} (k_i(\mathbf{c}) + k_j(\mathbf{c})) c_{ij}} \\ &= \sum_{\mathbf{k}} \sum_{\mathbf{c}} \delta_{\mathbf{k}, \mathbf{k}(\mathbf{c})} e^{\sum_{i < j} c_{ij} (2\alpha + \beta(k_i + k_j))} \\ &= \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} \sum_{\mathbf{c}} e^{\sum_{i < j} c_{ij} (2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j))} \\ &= \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} \prod_{i < j} (1 + e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}) \end{aligned} \quad (27)$$

where  $\delta_{\mathbf{k}, \mathbf{k}(\mathbf{c})} = \prod_i \delta_{k_i, k_i(\mathbf{c})}$  with  $\delta_{x,y}$  being the Kronecher delta, taking value 1 for  $x = y$  and 0 otherwise, and the Fourier representation of the Kronecher delta has been used

$$\delta_{x,y} = \int_{-\pi}^{\pi} \frac{d\Omega}{2\pi} e^{i\Omega(x-y)}. \quad (28)$$

Next, we focus on the normalised logarithm of (27), giving the free energy density  $f = -N^{-1} \log Z$ , which can be calculated exactly, for large  $N$ , in the finite connectivity regime  $\langle c_{ij} \rangle = \mathcal{O}(N^{-1}) \forall i, j$ , by using path integrals.

As a first step, we consider the likelihood  $\langle c_{ij} \rangle$  for two nodes  $i, j$  to be connected. This follows from the estimate of the average number of links  $L = \langle L(\mathbf{c}) \rangle = \sum_{i < j} \langle c_{ij} \rangle$ . Using

$$p(\mathbf{c}) = \frac{1}{Z} \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} \prod_{i < j} [e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)} \delta_{c_{ij}, 1} + \delta_{c_{ij}, 0}] \quad (29)$$

we have

$$\begin{aligned} L &= \sum_{i < j} \langle c_{ij} \rangle = \frac{1}{Z} \sum_{i < j} \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} \frac{e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}}{1 + e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}} \\ &\quad \times \prod_{k < \ell} [1 + e^{2\alpha + \beta(k_k + k_\ell) - i(\Omega_k + \Omega_\ell)}] \end{aligned}$$

$$= \sum_{i < j} \left\langle \frac{e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}}{1 + e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}} \right\rangle_{\mathbf{k}, \boldsymbol{\Omega}} \quad (30)$$

where

$$\langle \cdot \rangle_{\mathbf{k}, \boldsymbol{\Omega}} = \frac{1}{Z} \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\boldsymbol{\Omega}}{(2\pi)^N} e^{i\boldsymbol{\Omega}\mathbf{k}} \cdot \prod_{k < \ell} \left[ 1 + e^{2\alpha + \beta(k_k + k_\ell) - i(\Omega_k + \Omega_\ell)} \right]. \quad (31)$$

Hence, one has

$$\langle c_{ij} \rangle = \left\langle \frac{e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}}{1 + e^{2\alpha + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)}} \right\rangle_{\mathbf{k}, \boldsymbol{\Omega}} \quad (32)$$

In the regime  $\langle c_{ij} \rangle = \mathcal{O}(N^{-1})$ , it is convenient to transform  $\alpha \rightarrow \hat{\alpha} - \frac{1}{2} \log(N/c)$ , with  $c = \mathcal{O}(N^0)$ , to make (32) explicitly  $\mathcal{O}(1/N)$ , so we get

$$\langle c_{ij} \rangle \simeq \frac{c}{N} \langle e^{2\hat{\alpha} + \beta(k_i + k_j) - i(\Omega_i + \Omega_j)} \rangle_{\mathbf{k}, \boldsymbol{\Omega}} \quad (33)$$

and

$$Z = \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\boldsymbol{\Omega}}{(2\pi)^N} e^{i\boldsymbol{\Omega}\mathbf{k}} \exp \left[ \frac{c}{N} \sum_{k < \ell} e^{2\hat{\alpha} + \beta(k_k + k_\ell) - i(\Omega_k + \Omega_\ell)} \right]. \quad (34)$$

It is also useful to express the network distribution  $p(\mathbf{c})$  in the scaled parameter  $\hat{\alpha}$

$$\begin{aligned} p(\mathbf{c}) &= \frac{1}{Z} e^{2\alpha \sum_{i < j} c_{ij} + \beta \sum_{i < j} c_{ij} (k_i(\mathbf{c}) + k_j(\mathbf{c}))} \\ &= \frac{1}{Z} \prod_{i < j} \left[ e^{2\alpha + \beta(k_i(\mathbf{c}) + k_j(\mathbf{c}))} \delta_{c_{ij}, 1} + \delta_{c_{ij}, 0} \right] \\ &= \frac{1}{Z} \prod_{i < j} \left[ \frac{c}{N} e^{2\hat{\alpha} + \beta(k_i(\mathbf{c}) + k_j(\mathbf{c}))} \delta_{c_{ij}, 1} + \delta_{c_{ij}, 0} \right]. \end{aligned} \quad (35)$$

Next we introduce the following order parameters

$$P(k, \Omega | \mathbf{k}, \boldsymbol{\Omega}) = \frac{1}{N} \sum_{r=1}^N \delta_{k, k_r} \delta(\Omega - \Omega_r) \quad (36)$$

and insert into (34) for each  $(k, \Omega)$  the following integral:

$$\begin{aligned} 1 &= \int dP(k, \Omega) \delta[P(k, \Omega) - P(k, \Omega | \mathbf{k}, \boldsymbol{\Omega})] \\ &= (N/2\pi) \int dP(k, \Omega) d\hat{P}(k, \Omega) e^{iN\hat{P}(k, \Omega)P(k, \Omega) - i \sum_{r=1}^N \delta_{k, k_r} \delta(\Omega - \Omega_r) \hat{P}(k, \Omega)} \end{aligned} \quad (37)$$

Discretizing  $\Omega$  in steps of size  $\Delta$  which is eventually sent to zero, we can then write the free-energy as the following path integral, with the short-hand  $\{dP d\hat{P}\} = \prod_{k, \Omega} [dP(k, \Omega) d\hat{P}(k, \Omega)/2\pi]$  and sums over  $\Omega$  transformed into integrals:

$$\begin{aligned} f &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log Z \\ &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log \int \{dP d\hat{P}\} e^{iN \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega P(k, \Omega) \hat{P}(k, \Omega) + N \log \sum_{k \geq 0} \int d\Omega e^{i\Omega k - i\hat{P}(k, \Omega)}} \\ &\quad \times e^{\frac{Nc}{2} \sum_{k, k' \geq 0} \int_{-\pi}^{\pi} d\Omega d\Omega' P(k, \Omega) P(k', \Omega') e^{2\hat{\alpha} + \beta(k + k') - i(\Omega + \Omega')}} \\ &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log \int \{dP d\hat{P}\} e^{-N\Phi[P, \hat{P}]} \end{aligned} \quad (38)$$



where

$$\begin{aligned} \Phi[P, \hat{P}] = & -i \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega P(k, \Omega) \hat{P}(k, \Omega) - \frac{c}{2} \sum_{k, k' \geq 0} \int d\Omega d\Omega' P(k, \Omega) P(k', \Omega') e^{2\hat{\alpha} + \beta(k+k') - i(\Omega + \Omega')} \\ & - \log \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\Omega k - i\hat{P}(k, \Omega)} \end{aligned} \quad (39)$$

For large  $N$ , we can evaluate the path integral in (38) by steepest descent

$$f = \min_{P, \hat{P}} \Phi[P, \hat{P}] \quad (40)$$

Extremizing the action  $\Phi$  over  $P, \hat{P}$  leads to the saddle-point equations

$$\begin{aligned} P(k, \Omega) &= \frac{e^{i\Omega k - i\hat{P}(k, \Omega)}}{\sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\Omega k - i\hat{P}(k, \Omega)}} \\ -i\hat{P}(k, \Omega) &= ce^{\hat{\alpha} + \beta k - i\Omega} \left( \sum_{k \geq 0} e^{\hat{\alpha} + \beta k} \int_{-\pi}^{\pi} d\Omega P(k, \Omega) e^{-i\Omega} \right) \end{aligned} \quad (41)$$

The equation above suggests to define

$$P(k) = \int_{-\pi}^{\pi} d\Omega P(k, \Omega) e^{-i\Omega} \quad (42)$$

and

$$\gamma(\hat{\alpha}, \beta) = \sum_{k \geq 0} e^{\hat{\alpha} + \beta k} P(k) \quad (43)$$

so that

$$-i\hat{P}(k, \Omega) = ce^{\hat{\alpha} + \beta k - i\Omega} \gamma(\hat{\alpha}, \beta) \quad (44)$$

Hence,

$$P(k, \Omega) = \frac{e^{i\Omega k + c\gamma e^{\hat{\alpha} + \beta k} e^{-i\Omega}}}{\sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\Omega k + c\gamma e^{\hat{\alpha} + \beta k} e^{-i\Omega}}} \quad (45)$$

and

$$P(k) = \frac{(c\gamma e^{\hat{\alpha} + \beta k})^{k-1}}{(k-1)!} \left[ \sum_{k \geq 0} \frac{(c\gamma e^{\hat{\alpha} + \beta k})^k}{k!} \right]^{-1} \theta\left(k - \frac{1}{2}\right) \quad (46)$$

where  $\theta(x)$  is the Heaviside step function, taking value 1 for  $x > 0$  and 0 for  $x < 0$ . Note that  $P(k)$  is not a distribution because it is not normalised to one, instead we have  $\sum_{k \geq 0} P(k) = \langle e^{\beta k} \rangle_{\gamma}$  with

$$\langle \cdot \rangle_{\gamma} = \frac{\sum_{k \geq 0} (c\gamma e^{\hat{\alpha} + \beta k})^k / k!}{\sum_{k \geq 0} (c\gamma e^{\hat{\alpha} + \beta k})^k / k!}. \quad (47)$$

The resulting free energy density is

$$f(\hat{\alpha}, \beta) = \frac{c\gamma^2(\hat{\alpha}, \beta)}{2} - \log \sum_{k \geq 0} \frac{[e^{\hat{\alpha} + \beta k} c\gamma(\hat{\alpha}, \beta)]^k}{k!} \quad (48)$$

where  $\gamma$  solves the self-consistency equation

$$c\gamma^2(\hat{\alpha}, \beta) = \langle k \rangle_{\gamma}. \quad (49)$$

Finally, the ensemble parameters  $\hat{\alpha}, \beta$ , have to be determined from the equations for the constraints which are found by taking the derivatives of  $f$  with respect to  $\alpha, \beta$  as

$$\langle k \rangle = -\frac{\partial f}{\partial \alpha} = -\frac{\partial f}{\partial \hat{\alpha}} = -c\gamma \frac{\partial \gamma}{\partial \hat{\alpha}} + \langle k \rangle_\gamma + \frac{1}{\gamma} \langle k \rangle_\gamma \frac{\partial \gamma}{\partial \hat{\alpha}} = \langle k \rangle_\gamma \quad (50)$$

and

$$\langle k^2 \rangle = -\frac{\partial f}{\partial \beta} = -c\gamma \frac{\partial \gamma}{\partial \beta} + \langle k^2 \rangle_\gamma + \frac{1}{\gamma} \langle k \rangle_\gamma \frac{\partial \gamma}{\partial \beta} = \langle k^2 \rangle_\gamma \quad (51)$$

where the last equality in (50) and (51) follows from the saddle-point equation (49). Combining (50) and (49) we have  $\gamma = \sqrt{\langle k \rangle / c}$ , yielding the below equations for the ensemble parameters:

$$\langle k \rangle = \frac{\sum_{k \geq 0} k (\sqrt{c \langle k \rangle} e^{\hat{\alpha} + \beta k})^k / k!}{\sum_{k \geq 0} (\sqrt{c \langle k \rangle} e^{\hat{\alpha} + \beta k})^k / k!} \quad (52)$$

$$\langle k^2 \rangle = \frac{\sum_{k \geq 0} k^2 (\sqrt{c \langle k \rangle} e^{\hat{\alpha} + \beta k})^k / k!}{\sum_{k \geq 0} (\sqrt{c \langle k \rangle} e^{\hat{\alpha} + \beta k})^k / k!} \quad (53)$$

Hence, we can finally write the free-energy density as

$$f(\langle k \rangle, \langle k^2 \rangle) = -\frac{c}{2} - \log \sum_{k \geq 0} e^{\alpha k + \beta k^2} g(k) \quad (54)$$

where  $\alpha$  and  $\beta$  solve

$$\langle k \rangle = \frac{\sum_{k \geq 0} k g(k) e^{\alpha k + \beta k^2}}{\sum_{k \geq 0} g(k) e^{\alpha k + \beta k^2}} \quad (55)$$

$$\langle k^2 \rangle = \frac{\sum_{k \geq 0} k^2 g(k) e^{\alpha k + \beta k^2}}{\sum_{k \geq 0} g(k) e^{\alpha k + \beta k^2}} \quad (56)$$

with

$$g(k) = \frac{(\sqrt{c \langle k \rangle})^k e^{-(\langle k \rangle + c)/2}}{k!}. \quad (57)$$

In conclusion, we have that for the finitely-connected two-star model, with constrained average connectivity  $\langle k \rangle$  and degree variance  $\langle k^2 \rangle$ , the network distribution is

$$p(\mathbf{c} | \langle k \rangle, \langle k^2 \rangle) = \frac{1}{Z} \prod_{i < j} \left[ \frac{c}{N} e^{2\alpha + \beta(k_i(\mathbf{c}) + k_j(\mathbf{c}))} \delta_{c_{ij}, 1} + \delta_{c_{ij}, 0} \right] \quad (58)$$

where  $\alpha$  and  $\beta$  are determined from (55) and (56). We note that for the choice  $c = \langle k \rangle$ ,  $g(k)$  becomes a Poissonian distribution with parameter  $\langle k \rangle$ .

### 3.1. Test for $\beta = 0$

First, we check the validity of our equations for the case  $\beta = 0$ , where we should get back to the ER graphs. For  $\beta = 0$  the equation for the constraint (55) gives

$$\langle k \rangle = c e^{2\alpha} \quad (59)$$

Substituting in (58) we get

$$p(\mathbf{c}) = \frac{1}{Z} \prod_{i < j} \left[ \frac{\langle k \rangle}{N} \delta_{c_{ij}, 1} + \delta_{c_{ij}, 0} \right] \quad (60)$$

with

$$Z = \sum_{\mathbf{c}} \prod_{i < j} \left[ \frac{\langle k \rangle}{N} \delta_{c_{ij},1} + \delta_{c_{ij},0} \right] = \prod_{i < j} \left[ 1 + \frac{\langle k \rangle}{N} \right]$$

yielding, for large  $N$ ,

$$p(\mathbf{c}) = \prod_{i < j} \left[ \frac{\langle k \rangle}{N} \delta_{c_{ij},1} + \left( 1 - \frac{\langle k \rangle}{N} \right) \delta_{c_{ij},0} \right] \quad (61)$$

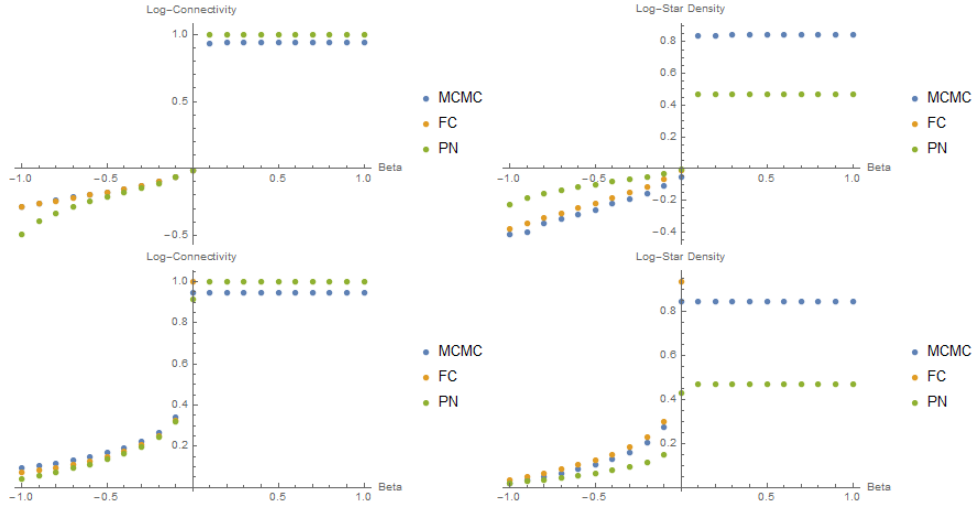
thus recovering the Erdős-Rényi ensemble.

### 3.2. Numerical results

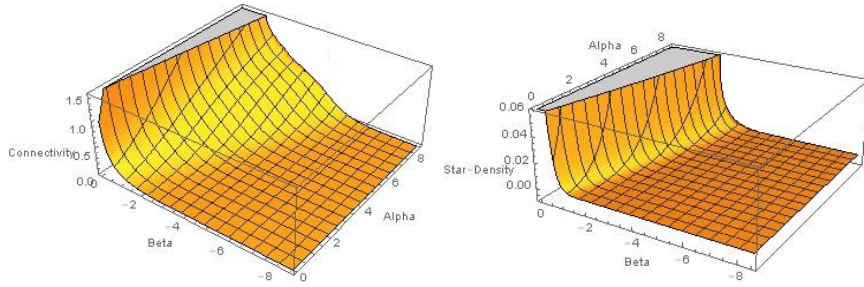
For  $\beta \neq 0$ , we need to solve equations (55), (56) numerically. We note, however, that for positive values of  $\beta$ , the sums on the RHS of these equations do not converge, hence only values  $\beta \leq 0$  are admissible. The parameter  $c$  can be chosen arbitrarily, so we will set it to unit, without loss of generality.

In Figure 1 we compare theoretical results from (55), (56) (orange symbols) with MCMC simulations for networks of  $N = 3000$  nodes (blue symbols) and predictions from mean-field theory (22) and (23) (green symbols). Plots show the logarithm of the link and of the star densities normalised with the logarithm of the system size, as functions of  $\beta$ , for fixed values of  $\alpha = -0.5, 4$ , corresponding to low and high connectivity respectively. MCMC simulations show a divergence in the links and stars densities for  $\beta > 0$ , consistently with the fact that equations (55), (56) do not converge in this regime, and show excellent agreement with theoretical predictions at  $\beta \leq 0$ . In particular, simulations data are on top of theoretical ones at low connectivity (top panel) and deviations stay within finite size effects at high connectivities (bottom panel). In contrast, mean-field predictions are seen to perform well at high connectivity but, as expected, get very inaccurate for small connectivity. In Figure 2, we show three dimensional plots of the average connectivity and average density of stars, as functions of  $\alpha$  and  $\beta$ . One has that for fixed values of  $\alpha$ , both connectivity and star density are at their highest for  $\beta = 0$  and they decay quickly for  $\beta < 0$ . Notably, the star density is always finite while the connectivity is finite, hence the model fails to produce an arbitrary number of stars for any given finite connectivity. In particular, we find that the star density is always close to its physical minimum. This is better understood by looking at contour plots of constant average connectivity and constant average star density in Figure 3. These show that for  $\beta$  fixed the connectivity increases with  $\alpha$  quicker than the star density, so that the star density decreases as we move along the contours of constant connectivity in the direction of decreasing  $\beta$ . Hence, the maximum number of stars is obtained for  $\beta = 0$ , which corresponds to Erdős-Rényi graphs, satisfying  $\langle k^2 \rangle - \langle k \rangle = \langle k \rangle^2$ , so that the contour of constant connectivity and star density coincide. The reason for this behavior can be understood by looking at equation (46), showing that for  $\beta \leq 0$ ,  $P(k)$  is Poissonian ( $\beta = 0$ ) or narrower ( $\beta < 0$ ), whereas for  $\beta > 0$ ,  $P(k)$  is not normalizable, with an asymptotic behavior for large- $k$  given by  $P(k) \sim e^{k(\beta k - \log k)}$ , independent of  $\alpha$ .

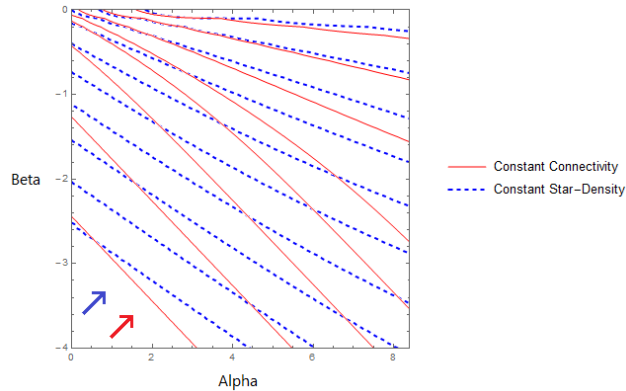
This shows that for  $\langle k^2 \rangle > \langle k \rangle^2 + \langle k \rangle$ , equations (55) and (56) have no solution. We will see in the next section that in this range of the imposed constraints, the partition function can no longer be calculated by the saddle-point method illustrated above and a different approach



**Figure 1.** Plot of  $\log\langle k \rangle / \log N$  (left panels) and  $\log\langle k^2 \rangle - \langle k \rangle / \log N$  (right panels) as functions of the ensemble parameter  $\beta$ , for  $\alpha = -0.5$  (top panels) and  $\alpha = 4$  (bottom panels) and  $c = 1$ . MCMC denote Monte Carlo simulations for  $N = 3000$ , FC denotes exact results from formulae (55), (56) and PN denotes predictions from expansions about mean-field theory (22), (23).



**Figure 2.** Plot of the average connectivity (left panel) and the average density of stars (right panel) as functions of the ensemble parameters  $\alpha, \beta$ .



**Figure 3.** Contour of constant average connectivity and average stars in the space of ensemble parameters  $\alpha, \beta$ .

is needed. This leads to a phase diagram in the  $\langle k \rangle - \langle k^2 \rangle$  plane with two phases which display different behaviors of the partition function  $Z$  in the large  $N$ -limit. The critical line separating the two phases is found by solving (55) and (56) for  $\beta = 0$ , which gives  $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$ . For  $\langle k^2 \rangle < \langle k \rangle^2 + \langle k \rangle$

$$Z(N\langle k \rangle, N\langle k^2 \rangle) \sim e^{-Nf(\langle k \rangle, \langle k^2 \rangle)} \quad (62)$$

where  $f(\langle k \rangle, \langle k^2 \rangle)$  is given by (54), whereas for  $\langle k^2 \rangle > \langle k \rangle^2 + \langle k \rangle$  the asymptotic behavior of  $Z$  is given by (69). We will show in the next section that the phase transition occurring at  $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$  is a condensation from a liquid to a condensed phase, due to a large deviation of sums of random variables. We note that condensation may be avoided in the scaling regime  $\beta = \mathcal{O}(\log N/N)$ , the same where Strauss model is known to display non-trivial behavior [11].

#### 4. Condensation as large deviation of sums of random degrees

In this section we show that the condensation transition occurring at  $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$  is related to a large deviation of sums of random degrees and we provide the asymptotics of  $Z$  in the condensate phase  $\langle k^2 \rangle > \langle k \rangle^2 + \langle k \rangle$ . To this purpose, it is convenient to use an alternative but equivalent definition of the two-star model, where the constraints on the links and stars are implemented directly in the network distribution

$$p(\mathbf{c} | \langle k \rangle, \langle k^2 \rangle) = \frac{1}{Z(N\langle k \rangle, N\langle k^2 \rangle)} \prod_{i < j} \left[ \frac{c}{N} \delta_{c_{ij}, 1} + \left( 1 - \frac{c}{N} \right) \right] \delta\left(\sum_i k_i(\mathbf{c}) - N\langle k \rangle\right) \delta\left(\sum_i k_i^2(\mathbf{c}) - N\langle k^2 \rangle\right) \quad (63)$$

via Kronecher deltas and links are drawn otherwise randomly and independently with likelihood  $c/N$ . By using path integrals and the saddle point method used in Section 3 we find that in the fluid phase  $\langle k^2 \rangle < \langle k \rangle^2 + \langle k \rangle$  the partition function  $Z(N\langle k \rangle, N\langle k^2 \rangle)$  has the large deviation behavior (62) with rate function given by the free-energy density (see Appendix B for full details)

$$f(\langle k \rangle, \langle k^2 \rangle) = \alpha \langle k \rangle + \beta \langle k^2 \rangle - \log \sum_k e^{\alpha k + \beta k^2} g(k) \quad (64)$$

where  $\alpha, \beta$  are determined from equations (55), (56) and  $g(k)$  is defined in (57). Up to additive constants, the free-energy density (64) is identical to (54), hence the definition (63) of the two-star model is thermodynamically equivalent to the standard definition (58). The marginal distribution  $p(k)$  in the fluid phase is given by (see Appendix C for derivations)

$$p(k) = g(k) \frac{e^{\alpha k + \beta k^2}}{\sum_{k \geq 0} g(k) e^{\alpha k + \beta k^2}} \quad (65)$$

showing that all random degrees contribute to the sums  $\sum_i k_i$ ,  $\sum_i k_i^2$  with small values.

Notably, the free-energy (64) and the marginal distribution (65) are identical to those of the system of independently and identically distributed random variables  $\mathbf{k} = \{k_1, \dots, k_N\}$  with "bare" distribution  $\hat{g}(k) = g(k)/\sum_k g(k)$  and global constraints on the average and the variance, introduced in [25]

$$p(\mathbf{k} | \langle k \rangle, \langle k^2 \rangle) = \frac{1}{Z(N\langle k \rangle, N\langle k^2 \rangle)} \prod_i \hat{g}(k_i) \delta\left(\sum_i k_i - N\langle k \rangle\right) \delta\left(\sum_i k_i^2 - N\langle k^2 \rangle\right) \quad (66)$$

The connection between the two models is transparent: a graph drawn from (63) will have a degree sequence  $\mathbf{k}$  where the degrees are random variables drawn independently from the Poissonian distribution  $\hat{g}(k)$  with average  $\sqrt{c\langle k \rangle}$ , subject to the two global constraints. For the choice  $c = \langle k \rangle$  each degree will be a Poissonian variable with average  $\langle k \rangle$ , subject to constraints.

Systems with factorised steady states (66) are known to display condensation for non-heavy-tailed distributions  $\hat{g}(k) \sim e^{-rk^\gamma}$  with  $1 \leq \gamma < 2$ , when conditioned on large deviations of their linear statistics, i.e. for  $\langle k^2 \rangle$  larger than a critical value  $\sigma(\langle k \rangle)$ , which is found by solving (55) and (56) for  $\beta = 0$  [25], i.e. from

$$\sigma(\langle k \rangle) = \frac{\sum_k k^2 g(k) e^{\alpha k}}{\sum_k g(k) e^{\alpha k}} \quad (67)$$

with  $\alpha$  solving

$$\langle k \rangle = \frac{\sum_k k g(k) e^{\alpha k}}{\sum_k g(k) e^{\alpha k}}. \quad (68)$$

For  $g(k)$  defined in (57), the critical value  $\sigma(\langle k \rangle)$  can be calculated exactly: equation (68) gives  $\alpha = \ln \sqrt{\langle k \rangle / c}$ , and substituting in (67) we obtain  $\sigma(\langle k \rangle) = \langle k \rangle^2 + \langle k \rangle$ .

For  $\langle k^2 \rangle > \sigma(\langle k \rangle)$ , the method developed in [25] yields for function (57) with asymptotics  $g(k) \sim e^{-k \log k}$ , the following asymptotic behaviour for the partition function

$$Z(N\langle k \rangle, N\langle k^2 \rangle) \sim e^{-NI(\langle k \rangle) - [N\langle k^2 \rangle - N\sigma(\langle k \rangle)]^{\frac{1}{2}} \log[N\langle k^2 \rangle - N\sigma(\langle k \rangle)]^{\frac{1}{2}}} \quad (69)$$

with

$$I(\langle k \rangle) = \alpha \langle k \rangle - \log \sum_{k \geq 0} \hat{g}(k) e^{\alpha k}$$

and  $\alpha$  determined from (68). In this regime, the marginal  $p(k)$  shows a bump at  $k \sim \sqrt{N}$ , meaning that there is a condensate of size  $\sim \sqrt{N}$  residing on a single site. Sampling networks from the two-star distribution (63) will then lead to graphs which have a bulk of homogeneous degrees and one single site accounting for all the degree heterogeneity in the network, making this model unsuitable as a null model for finitely connected random graphs with constrained average degree and variance. In addition, although Markov processes generating random variables  $\mathbf{k}$  under the two global constraints can be constructed as in [25], the definition of algorithms to sample networks  $\mathbf{c}$  from the measure (63) with the two hard constraints on degree average and variance, in an efficient and unbiased way, poses a challenge in its own right. We note that removing the hard constraint on the variance we obtain the system with factorised states studied in [18]

$$p(\mathbf{k} | \langle k \rangle) = \frac{1}{Z(N\langle k \rangle)} \prod_i \hat{g}(k_i) \delta(\sum_i k_i(\mathbf{c}) - N\langle k \rangle) \quad (70)$$

which is known to exhibit condensation only for heavy-tailed bare distributions  $\hat{g}(k) \sim Ak^{-\gamma}$ , with  $\gamma > 2$ , for  $\langle k \rangle < \sum_{k \geq 0} k \hat{g}(k)$ . For these models, the "dressed" marginal distribution is found to be [18]

$$p(k) = \frac{\hat{g}(k) e^{\alpha k}}{\sum \hat{g}(k) e^{\alpha k}} \quad (71)$$

with  $\alpha$  solving  $\langle k \rangle = \sum_k k p(k)$ . This suggests that condensation transitions in the two-star model may be avoided by removing the hard constraint on the variance and choosing the bare distribution

$\hat{g}(k)$  in such a way that the "dressed" marginal displays the desired variance  $\langle k^2 \rangle = \sum_k p(k)k^2$  while satisfying  $\sum_{k \geq 0} k\hat{g}(k) < \langle k \rangle$ .

## 5. Conclusion

In this work we analysed the finitely connected 2-star model. This model has been solved analytically within mean-field approximations, which predict a second-order phase transition between a symmetric and a symmetry-broken phase and a first-order transition in the link density occurring along the critical line separating the symmetry-broken phases, in the ensemble parameter space. In this work we solved the 2-star model exactly, in the thermodynamic limit, in the finite connectivity regime, where mean-field approximations are shown to become inexact. Our results show that in the thermodynamic limit the system undergoes a condensation transition, from a liquid to a condensed phase, related to large deviations in the sum of the random degrees, induced by the global constraints on their linear statistics. We showed that in the liquid phase, the degree statistics exhibited by the model is always in between regular and Erdős-Rényi graphs. Our results are in excellent agreement with MCMC simulations and are compared with existing results from mean-field theory and expansions around it. The latter become inaccurate for small connectivity, albeit providing the correct location of the critical line. When the global constraints imposed on the linear statistics insist on a degree statistics more heterogeneous than Erdős-Rényi graphs, the model undergoes a condensation transition, whereby a condensate residing on a single site appears in the system, which accounts for all the degree heterogeneity of the network, while the bulk of the network have homogeneous degrees.

We conclude that the finitely connected 2-star model is unsuitable, in its standard definition, as a null-model for real networks, with prescribed connectivity and link density. We suggest possible modifications of the model which may lead to non-trivial degree statistics in the finite connectivity regime, which either entail a different scaling of the Lagrange multiplier  $\beta$  constraining the stars or the use of a soft constraint for the star density, while the link density is hardly constrained.

## 6. Acknowledgments

AA wishes to thank Peter Sollich for useful discussions.

## 7. References

- [1] Caimo A and Friel N. Bayesian inference for exponential random graph models. *Social Networks*, 33:41–55, 2011.
- [2] Caimo A and Friel N. Bayesian model selection for exponential random graph models. *Social Networks*, 35:11–24, 2013.
- [3] Sanil A, Banks D, and Carley K. Models for evolving fixed node networks: Model fitting and model testing. *Social Networks*, 17(1):65–81, (1995).
- [4] Bollobás B. A probabilistic proof of an asymptotic formula for the number of labeled regular graphs. *Europ. J. Combinatorics*, 1(4):311–316, (1980).

- [5] Anderson C J, Wasserman S, and Croach B. A  $p^*$  primer: logit models for social networks. *Social Networks*, 21:37–66, (2002).
- [6] Banks D and Carley K M. Models of social network evolution. *J. Math. Sociology*, 21(1-2):173–196, (1996).
- [7] Strauss D and Ikeda M. Pseudolikelihood estimation for social networks. *J. Amer. Statistical Assoc.*, 85(409):204–212, (1990).
- [8] Bender E and Canfield E. The asymptotic number of labelled graphs with given degree sequences. *J. Combin. Theory Ser. A*, 24:296, (1978).
- [9] Bizhani G, Paczuski M, and Grassberger P. Statistical mechanics of networks. *Phys. Rev. E*, 86:011128, (2012).
- [10] Austad H and Friel N. Deterministic bayesian inference for the  $p^*$  model. *AISTATS*, pages 41–48, 2010.
- [11] Jonasson J. The random triangle model. *J. Appl. Prob.*, 36:852–867, (1999).
- [12] Park J and Newman M E J. Solution of the 2-star model of a network. *Phys. Rev. E*, 70(066146), 2004.
- [13] Park J and Newman M E J. Statistical mechanics of networks. *Phys. Rev. E*, 70:066117, (2004).
- [14] Park J and Newman M E J. Solution for the properties of a clustered network. *Phys. Rev. E*, 72:026136, (2005).
- [15] Besag J E. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, (1975).
- [16] Molloy M and Reed B. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161, (1995).
- [17] Handcock M S. *Statistical models for social networks: degeneracy and inference*, pages 229–240. National Academic Press, Washington, DC, (2003).
- [18] Evans MR, Majumdar SN, and Zia RKP. Canonical analysis of condensation in factorised steady states. *J. Stat. Phys.*, 123:357–390, (2006).
- [19] Frank O and Strauss D. Markov graphs. *J. Amer. Statistical Assoc.*, 81(395):832–842, (1986).
- [20] Erdős P and Rényi A. On random graphs i. *Publ. Math.*, 6:290, (1959).
- [21] Pattison P E and Robins G L. Neighbourhood-based models for social network. *Sociological Methodology*, 32:301–337, (2002).
- [22] Holland P W and Leinhardt S. An exponential family of probability distributions for directed graphs. *J. Amer. Statistical Assoc.*, 76(373), (1981).
- [23] Wasserman S and Robins G L. An introduction to random graphs, dependence graphs, and  $p^*$ . In *Models and Methods in Social Network Analysis*, pages 401–425. Cambridge University Press, (2005).
- [24] Wasserman S and Pattison P. Logit models and logistic regression for social networks: and introduction to markov graphs and  $p^*$ . *Psychometrika*, 61(3):401–425, (1996).
- [25] J Szavits-Nossan, Evans MR, and Majumdar SN. Condensation transition in joint large deviations of linear statistics. *J. Phys. A: Math. Theor.*, 47:455004, (2014).
- [26] Pattison T A B, Snijders amd P E, Robins G L, and Handcock M S. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, (2006).
- [27] Snijders T A B. The statistical evaluation of social network dynamics. In *Sociological Methodology*, Boston and London, (2001). Basil Blackwell.
- [28] Snijders T A B. Markov chain monte carlo estimation of exponential random graph model. *Journal of Social Structure*, 3(2), (2002).

## Appendix A. Mean-field predictions in the finitely connected regime

In the low connectivity regime we demand that  $p = c/N$  with  $c = \mathcal{O}(N^0)$ . To ensure that, we see from (19) that we have to set  $\beta = \mathcal{O}(1)$  and scale the ensemble parameter  $\alpha$  as  $\alpha \rightarrow \hat{\alpha} - 1/2 \log N$  so that the RHS of (19) becomes

$$\frac{1}{2} \left[ \tanh(\beta c + \hat{\alpha} - \frac{1}{2} \log N) + 1 \right] = \frac{1}{2} \left[ \frac{\frac{1}{N} e^{2\beta c + 2\hat{\alpha}} - 1}{\frac{1}{N} e^{2\beta c + 2\hat{\alpha}} + 1} + 1 \right] = \frac{\frac{1}{N} e^{2\beta c + 2\hat{\alpha}}}{\frac{1}{N} e^{2\beta c + 2\hat{\alpha}} + 1}$$



$$\sim \frac{1}{N} e^{2\beta c + 2\hat{\alpha}} \quad (\text{A.1})$$

where the last equality holds for large  $N$  and yields  $p = \mathcal{O}(N^{-1})$  as required. Equating this to the LHS of (19) we get the self-consistency equation

$$c = e^{2\beta c + 2\hat{\alpha}} \quad (\text{A.2})$$

A graphical analysis of (A.2) reveals that for  $\beta \leq 0$  there is only one solution, whereas for  $\beta > 0$  there can be one, two or no solutions depending on the choice of  $\hat{\alpha}$  and  $\beta$ . To identify the regions of the parameters for which solutions exist, note that the RHS of (A.2) is convex and strictly increasing in  $c$ . Let  $c^*$  be the value of  $c$  for which  $\frac{\partial}{\partial c} \text{RHS} = 2\beta e^{2\beta c + 2\hat{\alpha}} = 1$ . If the value of the RHS at  $c^*$  is greater than  $c^*$  then there can be no solutions of the self-consistency equation, while if it is less than  $c^*$  there are two, and there is only one solution if the RHS equals  $c^*$ . We find that  $c^*$  is given by

$$2\beta e^{2\beta c^* + 2\hat{\alpha}} = 1 \quad (\text{A.3})$$

We have that there exists only one solution when  $c^* = e^{2\beta c^* + 2\hat{\alpha}}$ , which using the defining equation of  $c^*$  (A.3) simplifies to  $c^* = 1/(2\beta)$ . Inserting in (A.3) we find that there is a unique solution for  $\beta = \frac{1}{2}e^{-1-2\hat{\alpha}}$ . We have no solutions when  $c^* < e^{2\beta c^* + 2\hat{\alpha}} = 1/(2\beta)$  which inserted in (A.3) gives  $1 < 2\beta e^{1+2\hat{\alpha}}$  hence  $\beta > \frac{1}{2}e^{-1-2\hat{\alpha}}$ . Finally we have two solutions for  $c^* > e^{2\beta c^* + 2\hat{\alpha}}$ , that is in the region  $\beta < \frac{1}{2}e^{-1-2\hat{\alpha}}$ . This implies that at  $\beta = e^{-1-2\hat{\alpha}}/2$  the only solutions of (A.2) is  $c = e^{1+2\hat{\alpha}}$ , whereas for  $\beta < e^{-1-2\hat{\alpha}}/2$  there are two solutions, one smaller and one greater than  $e^{1+2\hat{\alpha}}$ . As  $\beta$  approaches zero the greater solution tends to infinity while the lower solution tends to  $e^{2\hat{\alpha}}$ . From (17) we have  $e^\lambda = 1/(1-p)$  which substituted in (14) gives for the free-energy density  $f = F/N = \ln(1-p)(N-1)/2 \simeq -c/2$ , for  $p = c/N$  and  $N \gg c$ , showing that the free-energy decreases as the connectivity increases. Hence, the stable solution for  $0 \leq \beta \leq e^{-1-2\hat{\alpha}}/2$  will be the one with higher connectivity. In conclusion, the mean-field theory in the finite connectivity regime predicts a critical line at  $\beta = 0$  where the average connectivity jumps from  $\mathcal{O}(1)$  to  $\mathcal{O}(N)$  values. Although the mean-field approximation is invalid in the finite connectivity regime, it picks up the correct location of the critical line  $\beta = 0$  found from the exact analysis.

## Appendix B. Calculation of the partition function in the liquid phase

In this section we calculate the normalising constant  $Z(N\langle k \rangle, N\langle k^2 \rangle)$  of the distribution (63). First off, we use the identity  $1 = \sum_{\mathbf{k}} \delta_{\mathbf{k}, \mathbf{k}(\mathbf{c})}$  to write

$$Z(\mathbf{c} | N\langle k \rangle, N\langle k^2 \rangle) = \sum_{\mathbf{k}} \sum_{\mathbf{c}} \delta_{\mathbf{k}, \mathbf{k}(\mathbf{c})} \prod_{i < j} \left[ \frac{c}{N} \delta_{c_{ij}, 1} + \left(1 - \frac{c}{N}\right) \right] \delta\left(\sum_i k_i(\mathbf{c}) - N\langle k \rangle\right) \delta\left(\sum_i k_i^2(\mathbf{c}) - N\langle k^2 \rangle\right) \quad (\text{B.1})$$

and the Fourier representation of the Kronecher deltas, giving

$$\begin{aligned} Z(\mathbf{c} | N\langle k \rangle, N\langle k^2 \rangle) &= \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N\langle k \rangle + i\omega' N\langle k^2 \rangle} \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} e^{-i\omega \sum_i k_i - i\omega' \sum_i k_i^2} \\ &\quad \times \sum_{\mathbf{c}} \prod_{i < j} \left[ \frac{c}{N} e^{-i(\Omega_i + \Omega_j)} \delta_{c_{ij}, 1} + \left(1 - \frac{c}{N}\right) \right] \end{aligned} \quad (\text{B.2})$$

We next introduce the following order parameters

$$P(\Omega|\mathbf{\Omega}) = \frac{1}{N} \sum_{r=1}^N \delta(\Omega - \Omega_r) \quad (\text{B.3})$$

and insert into (B.2) for each  $\Omega$  the following integral:

$$\begin{aligned} 1 &= \int dP(\Omega) \delta[P(\Omega) - P(\Omega|\mathbf{\Omega})] \\ &= (N/2\pi) \int dP(\Omega) d\hat{P}(\Omega) e^{iN\hat{P}(\Omega)P(\Omega) - i \sum_{r=1}^N \delta(\Omega - \Omega_r)} \end{aligned} \quad (\text{B.4})$$

Discretizing  $\Omega$  in steps of size  $\Delta$  which is eventually sent to zero, and proceeding as in Sec. 3, we can then write the free-energy density as the path integral

$$\begin{aligned} f &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N \langle k \rangle + i\omega' N \langle k^2 \rangle} \\ &\quad \times \int \{dP d\hat{P}\} e^{iN \int d\Omega P(\Omega) \hat{P}(\Omega) + \frac{cN}{2} \left[ \int d\Omega d\Omega' P(\Omega) P(\Omega') e^{-i(\Omega + \Omega')} - 1 \right]} \\ &\quad \times \prod_i \sum_{k_i} \int \frac{d\Omega_i}{2\pi} e^{i\Omega_i k_i - i\omega k_i - i\omega' k_i^2 - i\hat{P}(\Omega)} \\ &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \int \{dP d\hat{P}\} e^{-N\Phi(P, \hat{P}, \omega, \omega')} \end{aligned} \quad (\text{B.5})$$

where

$$\begin{aligned} \Phi(P, \hat{P}, \omega, \omega') &= -i\omega \langle k \rangle - i\omega' \langle k^2 \rangle - i \int_{-\pi}^{\pi} d\Omega P(\Omega) \hat{P}(\Omega) - \frac{c}{2} \int d\Omega d\Omega' P(\Omega) P(\Omega') e^{-i(\Omega + \Omega')} \\ &\quad - \log \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\Omega k - i\omega k - i\omega' k^2 - i\hat{P}(\Omega)} + \frac{c}{2} \end{aligned} \quad (\text{B.6})$$

For large  $N$ , we can evaluate the integral by steepest descent

$$f = \min_{P, \hat{P}, \omega, \omega'} \Phi(P, \hat{P}, \omega, \omega') \quad (\text{B.7})$$

Extremizing the action  $\Phi$  over  $P, \hat{P}$  we obtain, proceeding as in Section 3,

$$c\gamma^2 = \langle k \rangle_{\gamma} \quad (\text{B.8})$$

with

$$\langle \cdot \rangle_{\gamma} = \frac{\sum_{k \geq 0} (c\gamma)^k e^{-i\omega k - i\omega' k^2} / k!}{\sum_{k \geq 0} (c\gamma)^k e^{-i\omega k - i\omega' k^2} / k!}. \quad (\text{B.9})$$

Extremizing  $\Phi$  over  $\omega, \omega'$  we obtain

$$\langle k \rangle = \langle k \rangle_{\gamma}, \quad \langle k^2 \rangle = \langle k^2 \rangle_{\gamma} \quad (\text{B.10})$$

Substituting the saddle point equations in the free energy we have

$$f = -i\omega \langle k \rangle - i\omega' \langle k^2 \rangle - \log \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{-i\omega k - i\omega' k^2} g(k) \quad (\text{B.11})$$

with

$$g(k) = \frac{(\sqrt{c\langle k \rangle})^k e^{-(c+\langle k \rangle)/2}}{k!} \quad (\text{B.12})$$

and  $\omega, \omega'$  determined from

$$\langle k \rangle = \frac{\sum_{k \geq 0} k g(k) e^{-i\omega k - i\omega' k^2}}{\sum_{k \geq 0} g(k) e^{-i\omega k - i\omega' k^2}} \quad (\text{B.13})$$

$$\langle k^2 \rangle = \frac{\sum_{k \geq 0} k^2 g(k) e^{-i\omega k - i\omega' k^2}}{\sum_{k \geq 0} g(k) e^{-i\omega k - i\omega' k^2}} \quad (\text{B.14})$$

Setting  $-i\omega = \alpha$ ,  $-i\omega' = \beta$  yields (64).

### Appendix C. Calculation of the marginal distribution in the liquid phase

In this section we calculate the marginal distribution

$$p(k|\mathbf{c}) = \frac{1}{N} \sum_i \delta_{k, k_i(\mathbf{c})} \quad (\text{C.1})$$

In the limit of large  $N$ , we expect this quantity to converge to its ensemble average

$$\begin{aligned} p(k) &= \langle p(k|\mathbf{c}) \rangle = \int \frac{dx}{2\pi} e^{ixk} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N \langle k \rangle + i\omega' N \langle k^2 \rangle} \\ &\quad \times \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} e^{-i\omega \sum_i k_i - i\omega' \sum_i k_i^2} \\ &\quad \times \sum_{\mathbf{c}} \prod_{k < \ell} \left[ \frac{c}{N} e^{-i(\Omega_k + \Omega_\ell) - ix(\delta_{i\ell} + \delta_{ik})} \delta_{c_{k\ell}, 1} + \left(1 - \frac{c}{N}\right) \delta_{c_{k\ell}, 0} \right] \\ &= \int \frac{dx}{2\pi} e^{ixk} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N \langle k \rangle + i\omega' N \langle k^2 \rangle} \\ &\quad \times \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} e^{-i\omega \sum_i k_i - i\omega' \sum_i k_i^2} \\ &\quad \times \exp \left\{ \sum_{k < \ell} \frac{c}{N} \left[ e^{-i(\Omega_k + \Omega_\ell) - ix(\delta_{i\ell} + \delta_{ik})} - 1 \right] \right\} \end{aligned} \quad (\text{C.2})$$

Next we work out the curly brackets as

$$\sum_{k < \ell} \frac{c}{N} \left[ e^{-i(\Omega_k + \Omega_\ell) - ix(\delta_{i\ell} + \delta_{ik})} - 1 \right] = \sum_{k < \ell} \frac{c}{N} \left\{ e^{-i(\Omega_k + \Omega_\ell)} [(e^{-ix} - 1)(\delta_{i\ell} + \delta_{ik}) + 1] - 1 \right\} \quad (\text{C.3})$$

Inserting the order parameter (B.3) via path integrals as done in Appendix B we arrive at

$$\begin{aligned} p(k) &= \int \frac{dx}{2\pi} e^{ixk} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N \langle k \rangle + i\omega' N \langle k^2 \rangle} \int \{dP d\hat{P}\} e^{iN \int d\Omega P(\Omega) \hat{P}(\Omega) + \frac{cN}{2} \left[ \int d\Omega d\Omega' P(\Omega) P(\Omega') e^{-i(\Omega + \Omega')} - 1 \right]} \\ &\quad \times \frac{1}{N} \sum_i \sum_{\mathbf{k}} \int \frac{d\Omega}{(2\pi)^N} e^{i\Omega \mathbf{k}} e^{-i\omega \sum_i k_i - i\omega' \sum_i k_i^2 - i \sum_i \hat{P}(\Omega_i) + c e^{-i\Omega_i} (e^{-ix} - 1) \int d\Omega P(\Omega) e^{-i\Omega}} \\ &= \int \frac{dx}{2\pi} e^{ixk} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N \langle k \rangle + i\omega' N \langle k^2 \rangle} \int \{dP d\hat{P}\} e^{iN \int d\Omega P(\Omega) \hat{P}(\Omega) + \frac{cN}{2} \left[ \int d\Omega d\Omega' P(\Omega) P(\Omega') e^{-i(\Omega + \Omega')} - 1 \right]} \\ &\quad \times \frac{1}{N} \sum_i \frac{\sum_{k_i} \int d\Omega_i e^{-i\Omega_i k_i - i\omega k_i - i\omega' k_i^2 - i \hat{P}(\Omega_i) + c e^{-i\Omega_i} (e^{-ix} - 1) \int d\Omega P(\Omega) e^{-i\Omega}}}{\sum_{k_i} \int d\Omega_i e^{-i\Omega_i k_i - i\omega k_i - i\omega' k_i^2 - i \hat{P}(\Omega_i)}} \end{aligned}$$

$$\begin{aligned}
& \times \prod_j \sum_{k_j} \int d\Omega_j e^{-i\Omega_j k_j - i\omega k_j - i\omega' \sum k_j^2 - i\hat{P}(\Omega_j)} \\
& = \int \frac{dx}{2\pi} e^{ixk} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \int \{dP d\hat{P}\} e^{-N\Phi([P, \hat{P}], \omega, \omega')} \\
& \times \frac{\sum_k \int d\Omega e^{i\Omega k - i\omega k - i\omega' k^2 - i\hat{P}(\Omega) + ce^{-i\Omega}(e^{-ix} - 1)} \int d\Omega P(\Omega) e^{-i\Omega}}{\sum_k \int d\Omega e^{i\Omega k - i\omega k - i\omega' k^2 - i\hat{P}(\Omega)}}
\end{aligned} \tag{C.4}$$

where  $\Phi(P, \hat{P}, \omega, \omega')$  is as in (B.6). We next calculate the integrals over  $\omega, \omega', P, \hat{P}$  by steepest descent. At the saddle point we have

$$-i\hat{P}(\Omega) = e^{-i\Omega} \sqrt{c\langle k \rangle} \tag{C.5}$$

and we obtain

$$p(k) = \int \frac{dx}{2\pi} e^{ixk} \frac{\sum_q \int d\Omega e^{i\Omega q - i\omega q - i\omega' q^2 + \sqrt{c\langle k \rangle} e^{-i\Omega} e^{-ix}}}{\sum_q \int d\Omega e^{i\Omega q - i\omega q - i\omega' q^2 + \sqrt{c\langle k \rangle} e^{-i\Omega}}} \tag{C.6}$$

where  $\omega, \omega'$  solve (B.13), (B.14). Performing the  $\Omega$ -integrals we have

$$p(k) = \int \frac{dx}{2\pi} e^{ixk} \frac{\sum_q e^{-i\omega q - i\omega' q^2} (ce^{-ix} \gamma)^q / q!}{\sum_q e^{-i\omega q - i\omega' q^2} (c\gamma)^q / q!} \tag{C.7}$$

and carrying out the integration over  $x$  finally gives

$$\begin{aligned}
p(k) &= \frac{e^{-i\omega k - i\omega' k^2} (\sqrt{c\langle k \rangle})^k / k!}{\sum_q e^{-i\omega q - i\omega' q^2} (\sqrt{c\langle k \rangle})^q / q!} \\
&= g(k) \frac{e^{-i\omega k - i\omega' k^2}}{\sum_k g(k) e^{-i\omega k - i\omega' k^2}}
\end{aligned} \tag{C.8}$$

with  $g(k)$  defined in (57). Setting  $-i\omega = \alpha$ ,  $-i\omega' = \beta$  finally yields (65).